Where in the World is this Image? Transformer-based Geo-localization in the Wild

Shraman Pramanick[1], Ewa M. Nowara[1], Joshua Gleason[2],
Carlos Castillo[1], Rama Chellappa[1]
[1]Johns Hopkins University, [2]University of Maryland, College Park

Introduction:

- **Geo-localization** is the task of predicting **geographic location** (latitude, longitude) from images.
- The goal of this work is **planet-scale geo-localization** from a **single image.**

Introduction:

- **Geo-localization** is the task of predicting **geographic location** (latitude, longitude) from images.
- The goal of this work is **planet-scale geo-localization** from a **single image.**

Challenges:

- Huge **diversity of scenes** all over the earth.
- **Appearance variation** of the same location depending on the **time of the day**, **weather, season**.

Where in the World is this Image? Transformer-based Geo-localization in the Wild

Introduction:

- **Geo-localization** is the task of predicting **geographic location** (latitude, longitude) from images.
- The goal of this work is **planet-scale geo-localization** from a **single image.**

Challenges:

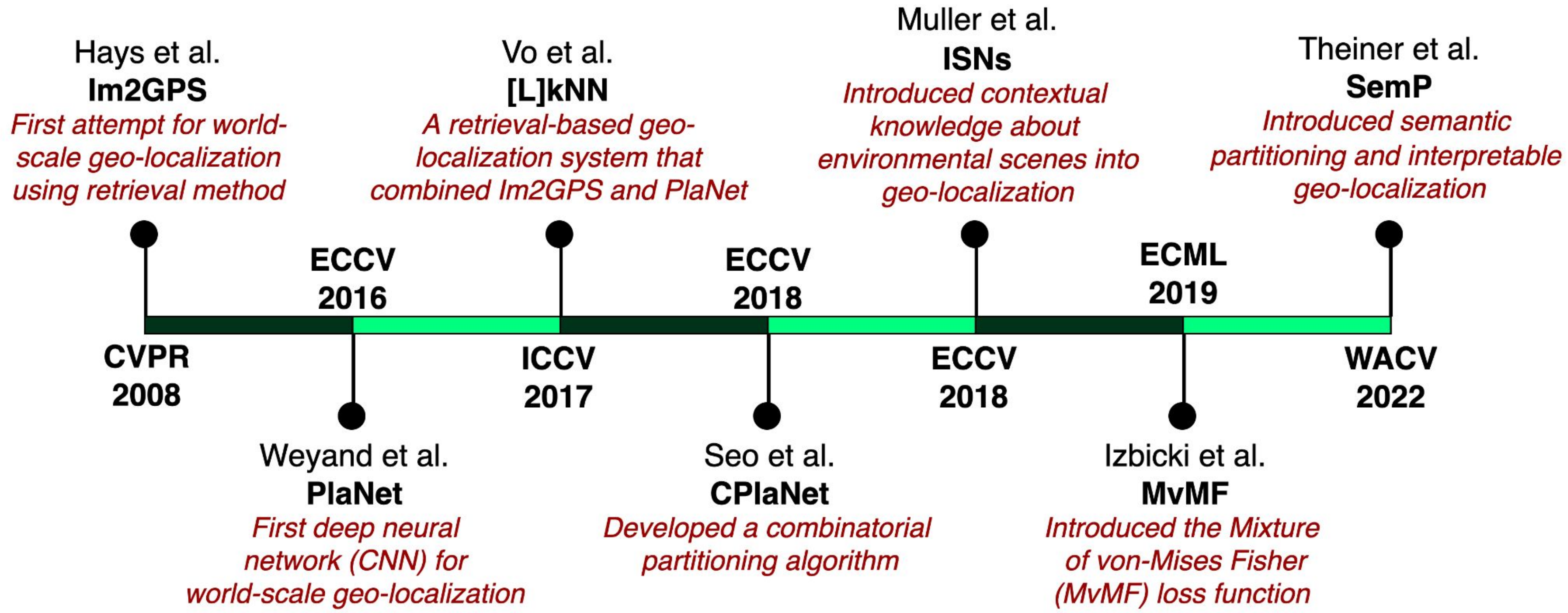
- Huge **diversity of scenes** all over the earth.
- **Appearance variation** of the same location under **different daytime or weather conditions**.

Where in the World is this Image? Transformer-based Geo-localization in the Wild

Related Works:

- **CNNs trained with large datasets** have significantly improved the performance of geo-localization methods and enabled extending the task to the scale of the entire world.

Planet-Scale Geo-localization Approaches:



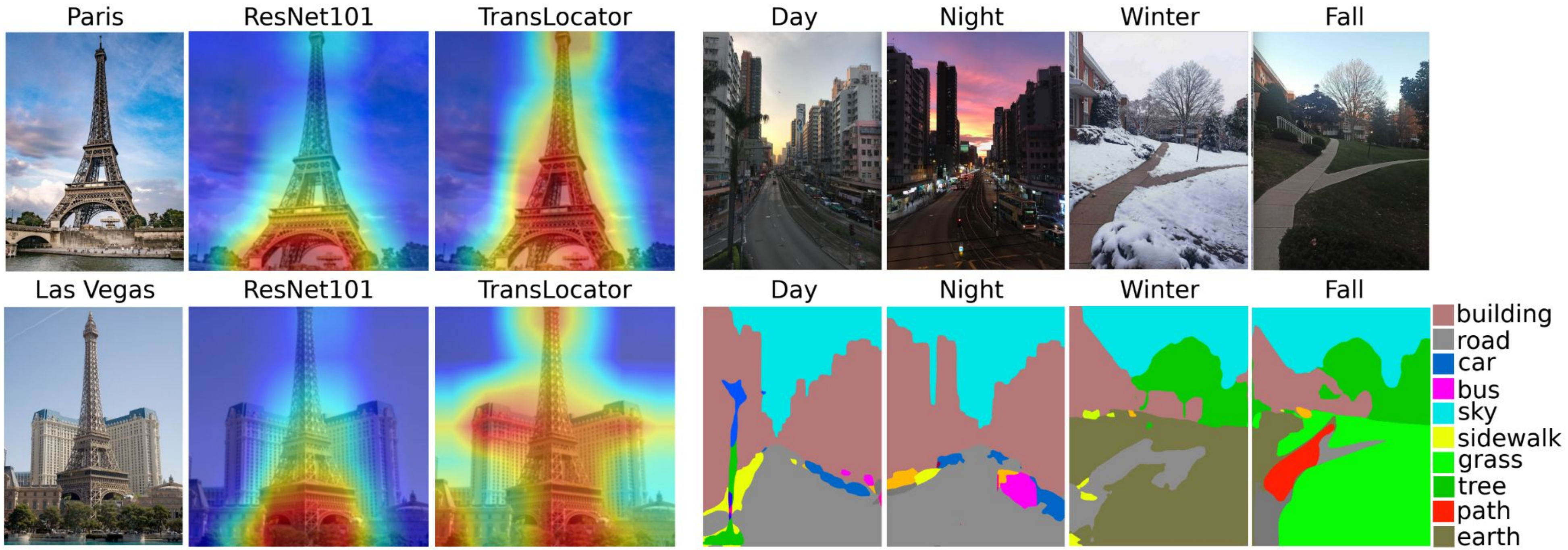
Approach:

- **Vision Transformer:** Early aggregation of global information helps to focus on fine-grained cues.
- **Semantic Segmentation:** Provides robustness to appearance variation at same location.
- **Multi-task Learning:** Predict the scene type (i.e., natural, urban, indoor) to better learn scene-specific features.

Where in the World is this Image? Transformer-based Geo-localization in the Wild

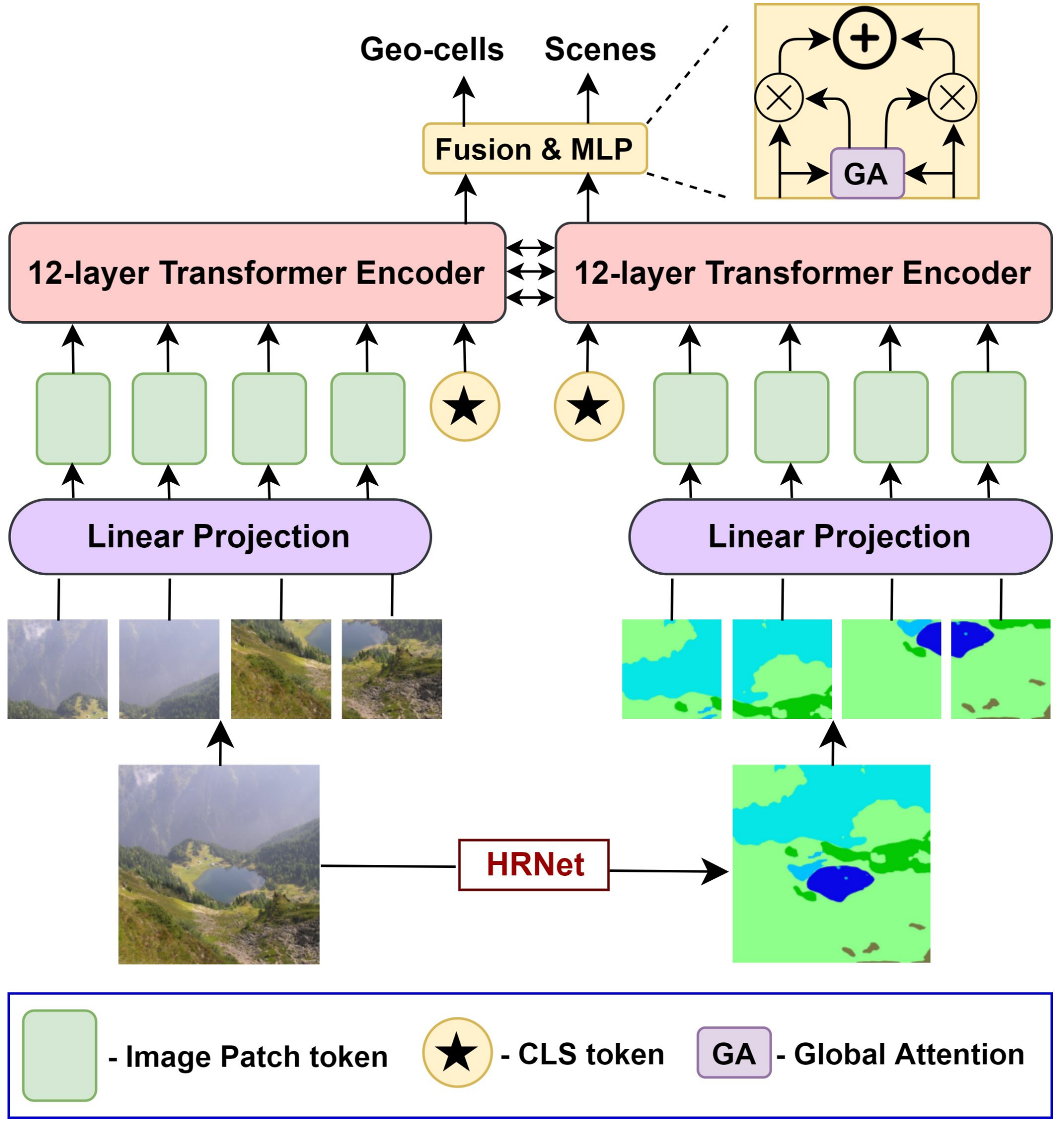
Approach:

- **Vision Transformer:** Early aggregation of global information helps to focus on fine-grained cues.
- **Semantic Segmentation:** Provides robustness to appearance variation at same location.
- **Multi-task Learning:** Predict the scene type (i.e., natural, urban, indoor) to better learn scene-specific features.



TransLocator:

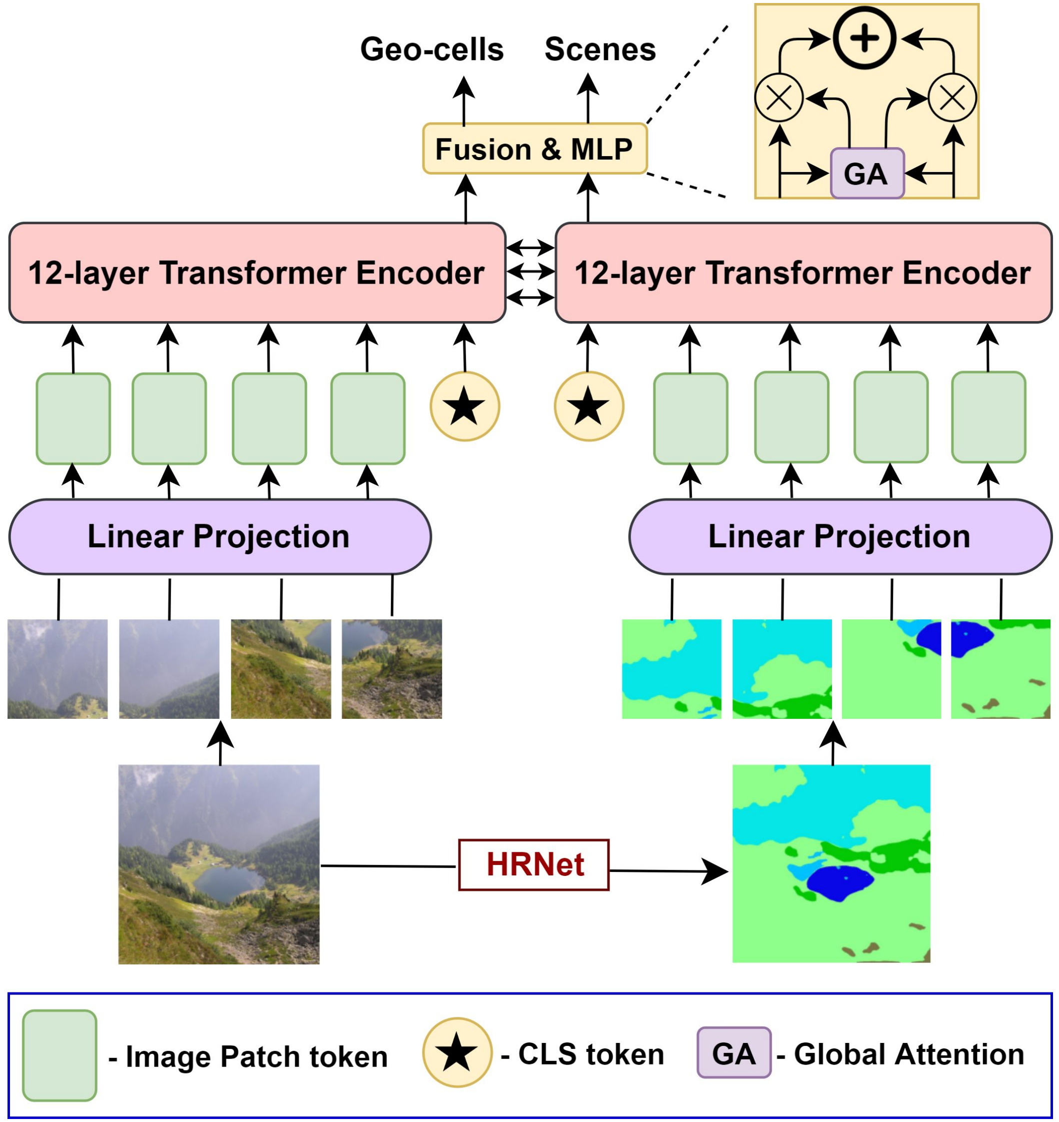
- **Dual-branch vision transformer** - RGB image and corresponding Semantic maps - complementary information of same input.



TransLocator:

- **Dual-branch vision transformer** - RGB image and corresponding Semantic maps - complementary information of same input.
- **Efficient and light-weight** fusion between two branches. We sum the CLS tokens of each branch after every transformer encoder layer.

$$ {}^{(i)}x^{(k)} = \left[g\left(\sum_{j \in \{\text{rgb, seg}\}} f({}^{(j)}x_{\text{cls}}^{(k)}) \right) || {}^{(i)}x_{patch}^{(k)} \right] $$



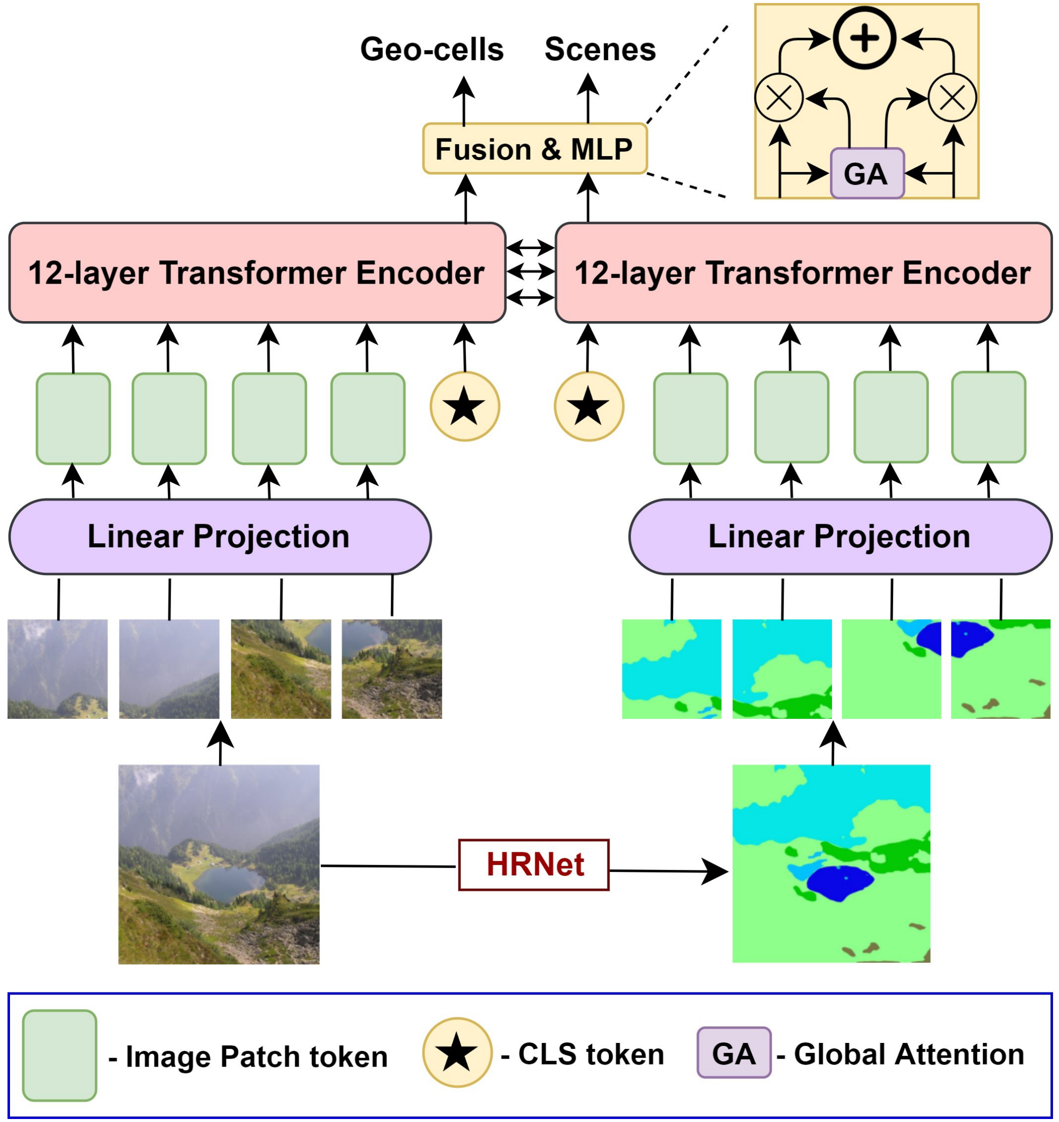
Where in the World is this Image? Transformer-based Geo-localization in the Wild

TransLocator:

- **Dual-branch vision transformer** - RGB image and corresponding Semantic maps - complementary information of same input.
- **Efficient and light-weight** fusion between two branches. We sum the CLS tokens of each branch after every transformer encoder layer.

$$^{(i)}x^{(k)} = \left[g\Big(\sum_{j \in \{\text{rgb, seg}\}} f\big(^{(j)}x_{\text{cls}}^{(k)}\big)\Big) \,||\, ^{(i)}x_{patch}^{(k)} \right]$$

- Different features are essential for various environmental settings, such as indoor and outdoor urban or natural scenes. **Geo-localization and scene recognition** are performed in **multi-task fashion**.



Experimental Results:

- **Dataset Used**
  - **Training:** MediaEval Placing Task 2016 dataset[1] (MP-16) containing 4.72M geo-tagged images sourced from Flickr.

[1]Larson, M. et al.; The benchmarking initiative for multimedia evaluation: Mediaeval 2016, IEEE MultiMedia, 2017

Experimental Results:

- **Dataset Used**
  - **Training:** MediaEval Placing Task 2016 dataset[1] (MP-16) containing 4.72M geo-tagged images sourced from Flickr.
  - **Validation:** YFCC26k[2], containing 25,600 geo-tagged images.

[1]Larson, M. et al.; The benchmarking initiative for multimedia evaluation: Mediaeval 2016, IEEE MultiMedia, 2017
[2]Theiner, J., et al.; Interpretable semantic photo geolocation, IEEE/CVF Winter Conference on Applications of Computer Vision, 2022

Experimental Results:

❖ **Dataset Used**

- ➢ **Training:** MediaEval Placing Task 2016 dataset[1] (MP-16) containing 4.72M geo-tagged images sourced from Flickr.
- ➢ **Validation:** YFCC26k[2], containing 25,600 geo-tagged images.
- ➢ **Evaluation:** Im2GPS[3], Im2GPS3k[4] and YFCC4k[5], containing 237, 2,997 and 4,536 geo-tagged images, respectively.

[1] Larson, M. et al.; The benchmarking initiative for multimedia evaluation: Mediaeval 2016, IEEE MultiMedia, 2017
[2] Theiner, J., et al.; Interpretable semantic photo geolocation, IEEE/CVF Winter Conference on Applications of Computer Vision, 2022
[3] Hays, J. et al.; Im2gps: estimating geographic information from a single image, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2008
[4] Hays, J. et al.; Large-scale image geolocalization, Multimodal Location Estimation of Videos and Images, Springer, 2015
[5] Vo, N. et al.; Revisiting im2gps in the deep learning era, IEEE/CVF International Conference on Computer Vision, 2017

Experimental Results:

- **Dataset Used**
  - **Training:** MediaEval Placing Task 2016 dataset[1] (MP-16) containing 4.72M geo-tagged images sourced from Flickr.
  - **Validation:** YFCC26k[2], containing 25,600 geo-tagged images.
  - **Evaluation:** Im2GPS[3], Im2GPS3k[4] and YFCC4k[5], containing 237, 2,997 and 4,536 geo-tagged images, respectively.
- **Reporting New State-of-the-Art Results**
  - Using TransLocator, we obtained the following continent-level performance improvements.
    - **Im2GPS: 5.5%, Im2GPS3k: 14.1%, YFCC4k: 4.9%, YFCC26k: 9.9%**

[1]Larson, M. et al.; The benchmarking initiative for multimedia evaluation: Mediaeval 2016, IEEE MultiMedia, 2017
[2]Theiner, J., et al.; Interpretable semantic photo geolocation, IEEE/CVF Winter Conference on Applications of Computer Vision, 2022
[3]Hays, J. et al.; Im2gps: estimating geographic information from a single image, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2008
[4]Hays, J. et al.; Large-scale image geolocalization, Multimodal Location Estimation of Videos and Images, Springer, 2015
[5]Vo, N. et al.; Revisiting im2gps in the deep learning era, IEEE/CVF International Conference on Computer Vision, 2017

Quantitative Results on Im2GPS:

Dataset	Method	Distance (a_r [%] @ km) Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS	Human	–	–	3.8	13.9	39.3
	[L]kNN, $\sigma = 4$	14.4	33.3	47.7	61.6	73.4
	MvMF	8.4	32.6	39.4	57.2	80.2
	PlaNet	8.4	24.5	37.6	53.6	71.3
	CPlaNet	16.5	37.1	46.4	62.0	78.5
	ISNs (M, f, S_3)	16.5	42.2	51.9	66.2	81.0
	ISNs (M,f*,S_3)	16.9	43.0	51.9	66.7	80.2
	ViT-MT	18.2	46.4	62.1	74.5	85.2
	TransLocator	**19.9**	**48.1**	**64.6**	**75.6**	**86.7**
	$\Delta_{\texttt{Ours - ISNs}}$	3.0 ↑	5.1 ↑	12.7 ↑	8.9 ↑	5.5 ↑

Ablation Experiments:

- **ViT-B/16** performs better than **ResNet101** and **EfficientNet-B4**.

Dataset	Method	Distance (a_r [%] @ km) Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS	ResNet101	14.3	41.4	51.9	64.1	78.9
	EfficientNet-B4	15.4	42.7	52.8	64.8	79.5
	ViT base	16.9	43.4	54.5	67.8	80.7
	+ Seg	17.6	44.8	58.9	70.0	83.3
	+ Seg + MFF	19.0	47.2	62.7	73.5	85.7
	+ Seg + MFF + Scene	**19.9**	**48.1**	**64.6**	**75.6**	**86.7**
Im2GPS 3k	ResNet101	9.0	25.1	32.8	46.1	63.5
	EfficientNet-B4	9.2	26.8	32.7	47.0	63.9
	ViT base	9.9	28.0	37.8	54.2	70.7
	+ Seg	10.5	29.1	42.5	55.8	73.6
	+ Seg + MFF	11.1	30.2	45.0	56.8	78.1
	+ Seg + MFF + Scene	**11.8**	**31.1**	**46.7**	**58.9**	**80.1**

Where in the World is this Image? Transformer-based Geo-localization in the Wild

Ablation Experiments:

- **ViT-B/16** performs better than **ResNet101** and **EfficientNet-B4**.
- Adding segmentation branch helps over single-branch (RGB) system. **Attention based fusion is better than concatenation-based fusion.**

Dataset	Method	Distance (a_r [%] @ km) Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS	ResNet101	14.3	41.4	51.9	64.1	78.9
	EfficientNet-B4	15.4	42.7	52.8	64.8	79.5
	ViT base	16.9	43.4	54.5	67.8	80.7
	+ Seg	17.6	44.8	58.9	70.0	83.3
	+ Seg + MFF	19.0	47.2	62.7	73.5	85.7
	+ Seg + MFF + Scene	**19.9**	**48.1**	**64.6**	**75.6**	**86.7**
Im2GPS 3k	ResNet101	9.0	25.1	32.8	46.1	63.5
	EfficientNet-B4	9.2	26.8	32.7	47.0	63.9
	ViT base	9.9	28.0	37.8	54.2	70.7
	+ Seg	10.5	29.1	42.5	55.8	73.6
	+ Seg + MFF	11.1	30.2	45.0	56.8	78.1
	+ Seg + MFF + Scene	**11.8**	**31.1**	**46.7**	**58.9**	**80.1**

Where in the World is this Image? Transformer-based Geo-localization in the Wild

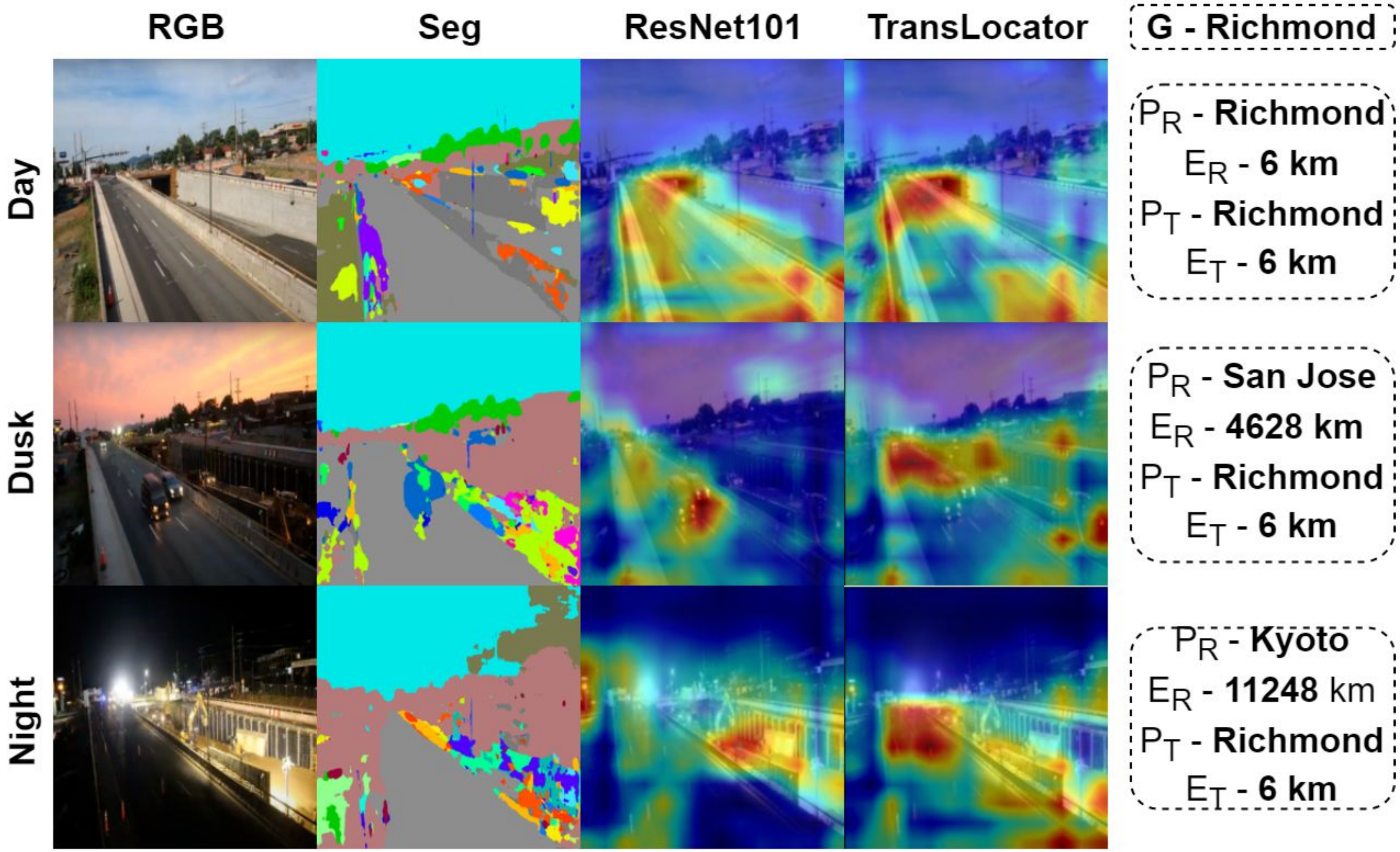
Ablation Experiments:

- **ViT-B/16** performs better than **ResNet101** and **EfficientNet-B4**.
- Adding segmentation branch helps over single-branch (RGB) system. **Attention based fusion is better than concatenation-based fusion.**
- Using multi-task learning further improves the performance.

Dataset	Method	Distance (a_r [%] @ km) Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS	ResNet101	14.3	41.4	51.9	64.1	78.9
	EfficientNet-B4	15.4	42.7	52.8	64.8	79.5
	ViT base	16.9	43.4	54.5	67.8	80.7
	+ Seg	17.6	44.8	58.9	70.0	83.3
	+ Seg + MFF	19.0	47.2	62.7	73.5	85.7
	+ Seg + MFF + Scene	**19.9**	**48.1**	**64.6**	**75.6**	**86.7**
Im2GPS 3k	ResNet101	9.0	25.1	32.8	46.1	63.5
	EfficientNet-B4	9.2	26.8	32.7	47.0	63.9
	ViT base	9.9	28.0	37.8	54.2	70.7
	+ Seg	10.5	29.1	42.5	55.8	73.6
	+ Seg + MFF	11.1	30.2	45.0	56.8	78.1
	+ Seg + MFF + Scene	**11.8**	**31.1**	**46.7**	**58.9**	**80.1**

Where in the World is this Image? Transformer-based Geo-localization in the Wild

Qualitative Results:



Where in the World is this Image? Transformer-based Geo-localization in the Wild

Error Analysis:

Examples of incorrectly localized Im2GPS images

G - Thailand
P - Morocco
Error - 10754 km

G - Hebei
P - Tokyo
Error - 2024 km

G - Alaska
P - Greenland
Error - 3936 km

G - Libya
P - Sudan
Error - 2019 km

Examples of incorrectly localized YFCC4k images

G - Varanasi
P - Agra
Error - 645 km

G - Jacksonville
P - West Mexico
Error - 2909 km

G - Colorado
P - Tokyo
Error - 9860 km

G - Berlin
P - San Jose
Error - 9138 km

Thanks for watching our presentation!

Sample Code and data is provided on Github:
https://github.com/ShramanPramanick/Transformer_Based_Geo-localization